CPSC 4300/6300 Applied Data Science

Simple Linear Regression

We will use the Boston data set for this tutorial, which contains housing values and other information for 506 neighborhoods around Boston. We will seek to predict median house value in $1000s (medv) using percent of households with low socioeconomic status (lstat) as predictor.

Open a new .R script in the RStudio editor and add author, date, and a brief description at the beginning of the file:

```
# Author: (Your Name)
# Date: September 14, 2018
# Purpose: Linear regression in R
```

Load the MASS and ISLR libraries, which contain data sets and functions we need:

```
library(MASS)
library(ISLR)
```

Look at the content of the Boston data set:

```
View(Boston)
names(Boston)
```

Question: What happens when you apply the summary() command to the data set?

Use attach to load the data set into the workspace. This way you can call columns in the data set directly without using the “$” operator:

```
attach(Boston)
```

Visually inspect the target (medv) and predictor (lstat) with tools of your choice.

1 Model fitting

Estimate a simple regression model with medv as target and lstat as predictor:

```
lm.fit <- lm(medv ~ lstat)
```

Type lm.fit into the R terminal to look at the regression output. Now apply summary() to lm.fit. How does the output differ?

Question: What is the association between median house value and percent of households with low socioeconomic status? Write down a sentence based on the output from the model.

To get a sense of the uncertainty associated with your estimated coefficients, estimate 95% confidence intervals:

```
confint(lm.fit)
```

2 Prediction

Now that we have trained the model, we can use the estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ to predict median house values for neighborhoods with different percentages of households with low socioeconomic status. Use the estimated coefficients and "by hand" (i.e., typing the correct equations into R) calculate predicted values for `medv` for the following values of `lstat`: 5, 10, and 15 percent.

Now use the `predict()` function with the same values for `lstat` as above. Also produce confidence intervals for your prediction:

```
predict(lm.fit, data.frame(lstat = c(5, 10, 15)), interval = "confidence")
```

Question: Compare the results to the prediction you have done "by hand". Do you get the same or different results?

Now produce prediction intervals instead of confidence intervals:

```
predict(lm.fit, data.frame(lstat = c(5, 10, 15)), interval = "prediction")
```

Question: How do the results differ? Why are some values different?

3 Visualizing results

Generate a scatterplot of the target and predictor along with the least squares regression line:

```
plot(lstat, medv)
abline(lm.fit)
```

Use `lwd` to adjust the width of the regression line and `col` to select a different color. Use the `pch` option to create different plotting symbols.

```
abline(lm.fit, lwd=3)
abline(lm.fit, lwd=3, col="red")
plot(lstat, medv, col="red")
plot(lstat, medv, pch=20)
plot(lstat, medv, pch="+")
```

Generate an overview of all available plotting symbols:

```
plot(1:20, 1:20, pch=1:20)
```

Next apply the `plot()` function to the regression output. This will produce four diagnostic plots:

```
plot(lm.fit)
```

To view all four plots together, use the `par()` function to divide the plotting region into a 2 × 2 panel grid:

```
par(mfrow=c(2,2))
plot(lm.fit)
```

4 Exercises

This exercise involves the use of simple linear regression on the `Auto` data set.

1. Use the `lm()` function to perform a simple linear regression with `mpg` (miles per gallon) as the response and `horsepower` (engine horsepower) as the predictor. Use the `summary()` function to print the results. Comment on the output. For example:

   (a) Is there a relationship between the predictor and the response?

   (b) How strong is the relationship between the predictor and the response?

   (c) Is the relationship between the predictor and the response positive or negative?

   (d) What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?

2. Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.

3. Use the `plot()` function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.